图书情报工作
第 64 卷 第 6 期 2020 年 3 月

学术论文引用预测研究进展*

■ 夏琬钧[1,2] 陈晓红[1] 江艳萍[1]

[1] 西南交通大学图书馆 成都 611756 [2]西南交通大学信息科学与技术学院 成都 611756

摘 要：[目的/意义]对学术论文引用预测影响因素和预测方法进行梳理，分析现存问题并提出发展方向。[方法/过程]采用文献调研法，综述国内外研究进展，总结预测影响因素和预测方法的相关内容和特点。[结果/结论]现有影响因素指标繁多，无统一标准；预测方法理论基础薄弱；引文预测动态性研究不足；预测模型通用性受限。未来应加强引文预测的理论研究、加强传统文献计量和替代计量的结合、加强自然语言处理的深度应用、建立统一的基线标准、构建更加精准的预测模型。

关键词：引用预测 影响因素 预测方法

分类号：G250

DOI：10.13266/j.issn.0252-3116.2020.06.016

1 引言

学术论文是传播科学知识的重要媒介，大多新技术、新发现都是通过学术论文进行公开。一项新成果往往是在前人工作基础上产生的，会引用他人文献，这也体现了对科学研究的继承和发展。由于科学研究的迅速发展使得每年都会产生大量学术论文，其数量呈指数级增长，研究人员如何从海量文献资源中快速找出有影响力的论文变得越来越具有挑战性，尤其是如何发现那些发表时间较短可能仅被少量文献引用但却代表最新研究成果的论文。当前，衡量论文影响力最简单、有效、客观的指标就是被引频次，人们普遍认为被高频次引用的论文反映了它对科学进步的贡献，因此，科学评价工作常常是根据论文被引频次进行的[1]。对被引频次进行预测不仅可以帮助研究者识别有参考价值的论文，而且有助于管理人员进行资源分配，是一项具有重要应用价值的任务。因此，本文将研究问题界定为预测每篇论文的被引情况，即从其他论文到该论文的引用，而非该论文引用其他论文[2]。目前这一主题已有不少研究成果，其中不乏综述性研究，如鲍玉芳等总结了常见引用预测方法[3]，但这些方法仅局限于回归分析，不够全面系统。鉴于此，本文采用文献调研法，对论文引用预测研究进行系统分析，重点关注近年来的最新进展，以期为今后工作提供借鉴和参考。

2 数据来源及分析

为了解国内外学术论文引用预测现状，笔者采用主题词“引用”（citation）和“预测”（prediction、predicting、predictor）对论文引用预测研究成果进行预检索。在此基础上，确定中文检索式“主题 =（论文 or 文献）and（引用 or 被引 or 引文 or 影响力）and 预测”和外文检索式“Title = citation * and predict *”，分别在CNKI 及 Web of Science 核心合集、SDOS、IEEE Xplore等数据库中进行检索，通过去重和人工判读，再利用参考文献数据进行扩展，基于最大相关度原则，共遴选出中文文献 25 篇，英文文献 112 篇。英文文献最早可追溯到 20 世纪 80 年代[4]，受到图书情报、计算机科学、生命科学、经济学等众多领域研究者的关注[5-8]。相比之下，中文文献数量不多且集中在近五年，可见中文文献的研究还处于起步阶段。通过进一步分析文献发现，现有引用预测方法主要是通过选择相关影响因素进行模型构建，从而实现引文预测。因此，本

* 本文系四川省文化和旅游厅图书情报学与文献学规划项目“基于学术大数据的潜力学者挖掘研究”（项目编号：WHTTSXM[2018]25）和四川省社会科学重点研究基地－四川学术成果分析与应用研究中心项目“基于在线评论的中文图书影响力研究”（项目编号：SCAA17－006）研究成果之一。

作者简介：夏琬钧（ORCID：0000－0001－9722－9837），馆员，博士研究生，E-mail：xiawanjun@home.swjtu.edu.cn；陈晓红（ORCID：0000－0003－3277－8725），副研究馆员，硕士；江艳萍（ORCID：0000－0003－3152－0204），馆员，博士。

收稿日期：2019－07－02 修回日期：2019－09－19 本文起止页码：138－145 本文责任编辑：杜杏叶

文从引用预测影响因素和具体预测方法两方面进行总结归纳。

3 引用预测影响因素研究现状

3.1 影响因素多元、开放

研究人员往往会关注影响论文被引的因素,以期增加自身成果的引用量[1]。而论文引用过程复杂,除了受到纯科学内容影响外,还受到其他因素影响,包括论文刊载期刊、作者声誉以及社会影响。早期学术论文引用预测影响因素的选择多从指标易获取性角度出发,较多地考虑论文、作者及期刊因素。随着研究不断深入,研究者开始在模型中融入其他可能影响引用预测的信息,如可以根据被引频次随时间动态变化规律、社交媒体数据等信息对论文未来被引情况进行预测,这也为影响因素相关研究提供了新的视角。目前,有关论文未来引用与影响因素关系的探讨已有大量卓有成效的工作,总体上可以将其归纳为论文因素、作者因素、期刊因素、时间因素和替代计量因素,具体指标及作用如表1所示:

表1 论文引用预测影响因素及作用

影响因素	具体指标	作用
论文因素	短期历史引文	短期历史引用数越多,未来被引概率越大[7]
	研究主题	热门话题通常会吸引更多关注和更多引文[5,9]
	参考文献(数量、年龄、影响力、多样性等)	参考文献数量和被引频次中度相关[10];参考文献平均年限越低,未来获得引用的可能性越大[11-12];具有高影响力参考文献的论文会得到更多引用[13];参考文献涉及领域越多样,后续被关注和被引用的概率越大[7]
	论文标题	带有娱乐性质的标题、复合标题或问题标题会吸引更多引用[14]
	论文长度	更长的文章可能与更详细的方法和结果有关,会增加科学工作的影响和传播,未来更有可能被引用[8]
	论文摘要及关键词	关键词在摘要中出现频率及关键词在期刊层面出现频率对未来引文数具有显著正相关作用[15]
	论文类型	综述性文章往往比研究型论文会得到更多引用[9]
	学科领域	跨学科性论文容易在未来获得更多引用[7]
作者因素	h指数及其衍生指数	高h指数作者撰写的论文更容易被引用[16]
	被引频次	之前被引频次高的作者容易获得更多引用[17]
	发文数量	作者发表论文数量越多,未来引文数就越高[7]
	研究领域	作者发文领域越多样,未来获得引用数越多[7]
	作者数量	论文作者人数与未来引用次数正相关[13]
	合作类型	国际合作论文与引文率呈显著正相关关系[13];跨学科合作会促进引文量的增加[8]
期刊因素	影响力(影响因子、被引频次)	论文刊载期刊影响力和论文未来引用正相关[18]
时间因素	发表时间	论文被引用概率随时间呈指数衰减[19]
	首次被引时间	首次被引越快的论文未来引文量越多[10]
	引文积累速率	引文积累速率越快的论文未来引文量越多[20]
替代计量因素	使用量(点击、下载、保存、浏览阅读等)	文献使用量、文献管理工具用户数和社交媒体提及量与未来引用次数之间存在着中等或显著正相关关系[21-25]
	文献管理工具用户数	
	社交媒体提及量	

3.2 影响因素研究特点

3.2.1 多维度

早期研究主要通过相关性分析考察未来引文数与个别因素的关系,但由于单类别因素包含信息有限,所以为了提高预测能力,研究人员试图综合更多维度因素,并对指标重要性进行分析。T. Chakraborty使用了包含论文、作者、期刊因素在内的16个指标构建预测模型,通过实证研究发现单独删除每个特征指标时,总体精度有不同程度下降[7],其中作者因素是最有效影响因素;耿骞等利用同样三大类因素共23个特征指标进行预测,发现单独使用其中某一类影响因素或者任意两类组合因素的预测效果比使用全部因素效果差,其中论文浏览下载次数、参考文献数量、作者被引次数在不同预测时间段内重要性排序都较为靠前[1];R. Yan考察了不同影响因素特征组的预测效果,研究发现论文、作者、期刊因素组合的预测效果最好,决定系数为0.927,单独使用某一类影响因素时,决定系数最高仅为0.659,其中作者影响力和期刊影响力是最为重要的指标[16]。

3.2.2 跨领域

目前大多数引文预测研究结论都不具有一般性,

因为使用的数据被限定在特定领域,所以解释影响文章引用率的因素大多只涉及单一学科,为了探寻是否存在通用性规律,研究者开始关注多学科应用。D. Wang 发表在 *Science* 上的研究成果推导出单个论文的引文动力学模型,发现不同学科和期刊的论文都倾向于遵循相同时间模式,表明共同的时间影响因素可以实现跨领域论文引用的长期预测[26];N. Onodera 选定了 6 个不同学科领域作为研究对象,发现了不同领域的一些共同影响因素指标,如近 5 年参考文献比率、参考文献数量等[27];F. Didegah 研究了 3 个不同学科领域引文影响因素,发现相同指标对不同学科领域的论文引文影响有较大差异,但同时也发现一些可以增加引文量的共同指标,如期刊影响力、参考文献影响力以及参考文献数量[13]。这些共性指标可以应用到不同学科领域中实现论文引用预测。

3.2.3 实时性

科研交流网络化极大地提高了科学传播效率,替代计量应运而生。替代计量数据由公用的 API 收集,数据开放,积累迅速[28],可以在一定程度上弥补传统文献计量指标时滞性缺陷。随着近年来替代计量各指标在学术数据应用中的不断发展完善,为论文引用预测增加了全新的影响因素,丰富了现有指标体系。熊泽泉等发现早期高下载和低下载论文更具预测性[21];H. Shema 研究表明科学博客引用的文章随后被引用次数往往比其他文章多[22];B. K. Peoples 发现 Twitter 推文数量与引用次数之间存在着很强的正相关性,比期刊影响因子更能预测引文率[23];D. Zollera 发现 BibSonomy 添加数、浏览数、导出数、查询数与未来引用次数之间存在中等相关性[24];M. Thelwall 对 Altmetic.com 的多种指标进行研究,发现 Mendeley 读者数量是未来引文影响的一致性预测指标[25]。

4 预测方法研究现状

4.1 预测方法多样、深入

随着科学计量学、网络科学、计算机科学的不断发展,学术论文引用预测产生了很多行之有效的方法,综合分析现有研究内容,可将其归纳为:统计学方法、机器学习方法、图模型方法。

4.1.1 统计学方法

统计学方法是论文引用预测早期最广泛使用的方法。它是通过对相关特征指标进行分析,获取统计数据,进而预测在未来一个时期的被引量。当前使用的统计学方法可以分为两类,具体如下:

(1)回归分析。为确定相关影响因素与未来被引量的因果联系,大多数学者采用回归分析法,如逐步回归、负二项回归、线性回归、分位数回归等[10,27,30-31]。T. Yu 利用逐步回归对图书情报学领域论文 5 年后影响力进行预测;C. Stegehuis 考虑到引文影响预测的高度不确定性,通过分位数回归预测论文发表后 5 年和 15 年被引频次的概率分布情况[31]。

(2)自定义模型。M. E. J. Newman 利用 Z 分数(计算一个时间窗口中发表论文的平均被引次数及其标准差,然后计算该论文被引次数与平均值之间的标准差)实现对物理学领域高被引论文的预测[32-33]。

4.1.2 机器学习方法

伴随人工智能技术的发展,机器学习算法在众多预测任务中表现优异。在学术论文引用预测方面,部分研究者将机器学习算法应用到学术大数据中,主要有三种方法,具体如下:

(1)分类。众多研究者将学术论文引用预测视为分类问题,因为此类模型有更好的泛化能力。研究者定义了诸多不同的分类标准,如 M. Wang[34]定义了三分类,即 3 年后的引文量是否为高被引、中被引或低被引;L. D. Fu 利用文章发表 10 年后引文量是否会超过 t(t = 20,50,100,500)进行分类[9];H. S. Bhat 根据引文分布百分位数(0,33%,66%)进行三分类[35]。另外,文中多采用 SVM、朴素贝叶斯、决策树、随机森林、AdaBoost、XGBoost 等算法进行分类预测,其中实验结果表明 SVM、随机森林、XGBoost 性能较高,准确率可以达到 90% 左右。

(2)聚类。X. Cao 等利用高斯混合模型(GMM)将具有相似引文模式的论文进行聚类,可以得到论文未来引文多个趋势及每种趋势的可能性大小[36],该方法简单有效,具有较强鲁棒性。

(3)回归。A. Abrishami 将引文预测看作一个回归学习问题,把递归神经网络(RNN)作为学习预测任务的强大模型,以此预测论文未来引文数量[37]。该方法仅通过引文数一个特征便得到较好的预测结果(决定系数最高可达 0.9),这是深度学习算法在论文引用预测方面的成功应用。

4.1.3 图模型方法

随着 PageRank 和 HITS 等网页排名算法的普及,基于图模型的方法被广泛应用于网络实体排序,在学术网络中,已有很多研究是通过引文和合著者关系对论文和研究者进行迭代排名[38]。在论文引用预测方面,图模型方法通过对论文"未来分值"进行计算实现

夏琬钧，陈晓红，江艳萍. 学术论文引用预测研究进展[J]. 图书情报工作,2020,64(6):138－145.

影响力排序，并根据被引频次排序进行准确性验证，从而间接实现引用预测。主要有两类方法，具体如下：

（1）简单图网络。N. Pobiedina 为了预测未来引文数量，将引文计数预测看作是引文网络中链路预测问题，基于频繁图模式挖掘引入 GERscore 分值来实现引文计数预测[39]；陈超美从科学图谱角度提出一种结构变异模型的预测性文献计量方法[40-41]。这些方法利用网络结构挖掘与引文预测相关的有用信息，但其结构较为单一，有可能忽略重要影响因素。

（2）复杂异构图网络。学术网络中包含了多种实体和关系类型，节点之间相互影响[42]，使得网络结构具有高度复杂性和异构性，从异构网络中可以挖掘出更多隐藏信息。FutureRank[43]是较早将异构网络用于论文未来引文排序的算法，该算法通过构建论文引用网络和作者论文网络，进行随机游走迭代计算，准确率可以达到 75%。后续研究多在此基础上进行优化改进，如刘大有等通过计算作者撰写权威值和引用权威值而不需要计算论文 PageRank 值，性能大幅度提升[44]；MRCoRank 算法构建了具有时间感知的加权网络，并在网络中融入文本信息，利用基于突发词检测方法可以预测到开创性论文[48]；NERank 算法将论文、作者、期刊三种类型节点表征到同一低维向量空间，同时考虑网络全局和局部结构信息，预测准确率可以比 MRCoRank 提升 6%[47]。

4.2 不同预测方法对比

三种主流预测方法对比分析见表 2。统计学方法主要是用来分析和理解数据，利于发现引文和各影响因素之间的关系及这些影响因素的重要性，可以得到严谨的数学解释和推理公式，以强大的数学理论支撑解释因果，在数据量有限的情况下，较易发现各影响因素之间的相关性。机器学习方法追求的是预测准确性，可以充分利用学术大数据中的高维度特征，对论文引用进行精准预测，而统计学方法在处理这种大数据、高维度特征问题时，收敛速度和预测精度都无法达到满意的效果。另外，虽然统计学和机器学习均包含回归方法，但二者不同，统计回归注重的是对历史数据的无偏差拟合，而机器学习回归则是减少方差尽量避免过拟合现象的发生，以获得更高的预测准确率。统计学和机器学习方法通常将所有引用视为“平等”，而图模型方法可以充分利用引用网络、作者网络等可用的结构信息，为引用赋予不同的权重以区分高质量和低质量引用或区分来自不同影响力学者的引用，可以更清晰地揭示论文被引趋势。

表 2　不同预测方法对比分析

方法	适用范围	评价指标	预测效果	优点	缺点
统计学	多用于探索影响因素和引文关系，适用小数据量	相关系数(R)、决定系数(R^2)、均方根误差(RSME)、均方残差(MSR)等	预测准确率通常低于90%	具有可解释性	容易产生过拟合，预测能力有限
机器学习	多用于大规模数据集，适合高维特征	AUC、F1 值、ROC 等	预测准确率最高可达90%	能够获得可重复预测的模型，准确率较高	缺乏可解释性
图模型	多用于引文细分研究	TOP N 排名准确度等	预测准确率通常低于90%	可以挖掘学术网络中的隐藏关系	计算复杂度高

4.3 预测方法研究特点

4.3.1 大数据

早期论文引用预测基本都是小样本统计分析，使用的数据规模仅有几百条，而大数据技术不断发展，改变了研究范式。随着数据规模的提升，研究者可以快速地从海量数据中获取有价值的信息，学术论文引用预测各方法中所用数据集从最初几百条发展到如今的几十万、几百万甚至上千万条，如 H. S. Bhat 将机器学习算法应用在一个包含 300 多万独立作者的近 800 万条论文记录的大型数据集上，用于存储数据的 JSON 文件达 220GB[35]。大数据分析有助于发现更多信息和规律，样本量的增多更加有助于构建机器学习模型。在很多情况下，所处理的数据规模越大，机器学习模型的效果越好。

4.3.2 智能化

机器学习使充分利用数据中蕴含的知识与价值来实现数据智能化处理成为可能[46]。基于机器学习的方法和单纯依靠人为设计模型的方法相比，可以避免很多主观性因素干扰，更加聚焦“数据”本身，从历史数据中自动地学习出规则，从而实现对新数据的预测。如 L. D. Fu 利用决策树算法对文献计量学特征进行学习，自动提取特征模式，减少人工干预[9]。同时，学者们开始考虑将预测算法应用到智能化检索系统中，如 R. Yan 等结合论文引用预测算法实现了论文个性化推

荐系统的原型设计[16]；沈雷设计和开发了论文影响力预测系统，并可通过移动端进行结果展示[47]。虽然这些智能化系统离全面应用还有一定差距，但为论文引用预测算法的真正落地提供了相关实践指导。

4.3.3 结构化

论文引用预测已不再是简单依靠单个影响因素，而是在结构化网络中综合考虑作者权威、期刊影响力等相互作用。在结构化网络中，每个论文节点通过引文链接到另一个论文节点，帮助我们获得更多关于作者、论文的信息，如 S. Wang 构建了论文、作者、期刊和文本特征的多个子网络，并利用各子网络之间的相互强化关系实现论文引用预测排序[48]；曾玮利用论文-论文引用关系矩阵、作者-文献关系矩阵构建预测模型，获得较高的执行效率[49]。利用结构化网络将已知信息带入到学术网络中，使节点之间产生"互动"，从而通过网络结构和拓扑性质增强实现对引用行为的理解。

5 存在问题和未来展望

5.1 存在问题

总的来说，学术论文引用预测研究得到了大量研究者的关注，各种新的预测方法丰富和拓展了研究内容，但同时，仍存在很多尚未得到解决的问题。

（1）影响因素指标繁多，无统一选择标准。研究表明输入特征选择是产生高效预测的原因[35]，但当前影响论文引用预测指标繁多，如何确定影响论文引用的主要因素仍然是一个复杂问题，尽管对这一问题已进行了大量研究，但尚未达成共识，甚至有相互矛盾的结论。这在一定程度上是因为现有研究主要集中在一个因素（或多个因素相互独立），并未充分考虑不同因素之间的相互作用；另一个原因是不同领域特征选择标准不同或产生的影响不同，尽管已有跨领域研究，但研究结论不尽相同。

（2）预测方法理论基础薄弱。现有方法大多是参数化的，需要对模型参数进行准确估计，才能做出正确预测，而引用动力学的复杂模式很难用一个简单参数模型来描述。同时，参数设置往往需要人工不断调整以确定最佳数值，所以目前大多数参数估计主观因素较多，缺乏深层次的理论研究，而所识别的影响未来论文引用的多维指标也缺乏科学判据，各指标和未来引文数量只有相关性联系，并无明确因果关系，无法解释可能影响引文的混杂因素，也不能很好解释预测行为及预测出错的原因。

（3）引文预测动态性研究不足。引文预测是一个动态变化过程，相同影响因素作用大小有可能随着时间推移产生变化。前一年的模型是否对下一年继续适用还不甚明确，不同预测时间窗内是否有不同影响因素起作用也需要进一步研究。另外，现有基于图模型的方法大多还是静态的，而网络是动态发展的，随着时间推移，网络中节点和链接的数量都在不断更新，尚缺乏对时序拓扑信息的充分研究，这些都会影响预测准确性。

（4）预测模型通用性受限。现有研究不能适用所有引文模式，虽然 D. Wang 提出了遵循相同时间模式的预测模型，但也同时指出了模型的局限性，无法适用于若干年后出现二次引文高峰的论文[26]。当前研究对预测时间的选择具有随机性，未充分考虑到学科文献半衰期的影响[3]。有些研究认为论文发表后 5 年的影响力是论文质量的重要体现，10 年后的引文预测并不重要，但对于数学、经济学等学科实际上可能需要较长的预测时间。同时，应用于不同的预测时间周期时，模型是否有效尚且未知。

5.2 未来展望

（1）加强对论文引用预测的理论研究。未来应重点加强对引用量驱动因素的理论探究，以便更好解释预测结果。同时，引用行为是一个动态变化过程，具有复杂的时间异质性。在进行模型构建时，可以借鉴更多的研究方法，如复杂网络、系统动力学和演化理论，加深对引文网络拓扑结构及网络演化的理解，把握其动态演化趋势和特征，探索引文动态变化规律，从而夯实理论研究基础。

（2）加强传统文献计量和替代计量的结合。学术交流形式的网络化为学术信息传播和讨论带来了新场所，替代计量学突破了传统计量学局限，具有实时性，但也存在数据源、计量指标可靠性不足等问题。未来研究应将传统计量指标和替代计量指标充分结合，利用替代计量指标的"补充"作用，探索在它们相互作用下对未来引文的影响，进行深入的相关性分析，构建更加有效的影响因素特征空间。

（3）加强自然语言处理的深度应用。将科学计量学理论知识和自然语言处理技术相结合，充分发挥自然语言处理在语义关联与挖掘方面的强大功能，探索

更加深层次的指标,从而避免"可操纵"因素影响,实现基于内容层面的分析。从学术文献全文数据中挖掘出更多内容特征知识,加入基于文本的引用内容分析,进行内容建模,揭示引用动机,客观判断文献价值,以提升预测准确率。

(4)建立统一的基线标准。构建统一基线标准有利于在前人的研究中继续进行科学发现。现有数据多样化,各个方法都采用不同数据集、不同影响因素进行研究,而引文预测的研究往往受到所用数据集的完备性和正确性限制,如果没有统一标准,设计的特征指标会缺乏健壮性和可理解性。如何构建一个统一的学术数据基准,需要研究者共同分享数据和指标,这也是未来论文影响力预测研究亟需解决的重要问题。

(5)构建更加精准的预测模型。未来研究应加强对不同引文模式的深入研究,如"睡美人"论文[50],以探索统一的预测框架;加强对引用行为的细分研究,如自引、消极引用或来自不同影响力学者的引用等,进一步研究作者社会关系、各种引文分布影响,考虑并加入更精准的时序信息,构建动态序列预测模型。对这些问题更为细粒度的研究可以充分地描述所观察到的现象,揭示现象背后的机制或过程, 以确立精确因果关系。

6 结语

综上所述,在智能化、数字化、网络化的环境下,学术论文引用预测研究内容不断更新,产生了新的影响因素指标和预测方法,本文对近年来的最新进展进行了梳理总结,但当前仍存在很多开放性问题需要解决,未来应深入理论研究,推进理论创新,加强新指标和新方法的合理运用,促进数据集共享,推动数据开放,将预测方法整合到智能学术搜索平台中,满足用户多元化、个性化需求,实现学术论文引用预测方法的科学应用。

参考文献:

[1] 耿骞,景然,靳健,等. 学术论文引用预测及影响因素分析[J]. 图书情报工作,2018,62(14):29－40.

[2] YANG L, ZHANG Z, CAI X, et al. Citation recommendation as edge prediction in heterogeneous bibliographic network: a network representation approach[J]. IEEE Access, 2019, 7: 23232－23239.

[3] 鲍玉芳,马建霞. 科学论文被引频次预测的现状分析与研究[J]. 情报杂志,2015,34(5):66－71.

[4] STEWART J A. Achievement and ascriptive processes in the recognition of scientific articles[J]. Socialforces,1983,62(1):166－189.

[5] WILLIS D L, BAHLER C D, NEUBERGER M M, et al. Predictors of citations in the urological literature[J]. BJUinternational,2011,107(12):1876－1880.

[6] KOSTEAS V D. Predicting long-run citation counts for articles in top economics journals[J]. Scientometrics,2018,115(3):1395－1412.

[7] CHAKRABORTY T,KUMAR S,GOYAL P,et al. Towards a stratified learning approach to predict future citation counts[C]//2014 IEEE/ACM joint conference on digital libraries (JCDL). London: IEEE computer society,2014:351－360.

[8] ANTONIOU G A,ANTONIOU S A,GEORGAKARAKOS E I,et al. Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature[J]. Annals of vascular surgery,2015,29(2):286－292.

[9] FU L D, ALIFERIS C F. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature[J]. Scientometrics,2010,85(1):257－270.

[10] YU T, YU G, LI P Y, et al. Citation impact prediction for scientific papers using stepwise regression analysis[J]. Scientometrics,2014,101(2):1233－1252.

[11] HASLAM N, BAN L, KAUFMANN L, et al. What makes an article influential? Predicting impact in social and personality psychology[J]. Scientometrics,2008,76(1):169－185.

[12] ROTHC , Wu J , Lozano S . Assessing impact and quality from local dynamics of citation networks[J]. Journal of informetrics,2013,6(1):111－120.

[13] DIDEGAH F, THELWALL M. Which factors help authors produce the highest impact research? Collaboration, journal and document properties[J]. Journal of informetrics,2013,7(4):861－873.

[14] SUBOTIC S , MUKHERJEE B . Short and amusing: the relationship between title characteristics, downloads, and citations in psychology articles[J]. Journal of information science, 2014, 40(1):115－124.

[15] SOHRABI B, IRAJ H. The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts[J]. Scientometrics,2017, 110(1):243－251.

[16] YAN R, TANG J, LIU X, et al. Citation count prediction: learning to estimate future citations for literature.[C]// ACM international conference on information & knowledge management. Glasgow:ACM,2011:1247－1252.

[17] TAHAMTAN I, AFSHAR A S, AHAMDZADEH K. Factors affecting number of citations: a comprehensive review of the literature[J]. Scientometrics,2016,107(3):1195－1225.

[18] BORNMANN L, LEYDESDORFF L, WANG J. How to improve

the prediction based on citation impact percentiles for years shortly after the publication date? [J]. Journal of informetrics, 2014, 8(1):175 –180.
[19] 张美平, 尚明生. 基于持续关注度衰减的重要论文预测[J]. 复杂系统与复杂性科学, 2015, 12(3):77 –84.
[20] BORNMANN L, DANIEL H D. Citation speed as a measure to predict the attention an article receives: An investigation of the validity of editorial decisions at AngewandteChemie International Edition[J]. Journal of informetrics, 2010, 4(1):83 –88.
[21] 熊泽泉,段宇锋. 论文早期下载量可否预测后期被引量?——以图书情报领域期刊为例[J]. 图书情报知识,2018(4):32 –42.
[22] SHEMA H, BAR-ILAN J, THELWALL M. Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics[J]. Journal of the Association for Information Science and Technology,2014,65(5):1018 –1027.
[23] PEOPLES B K, MIDWAY S R, SACKETT D, et al. Twitter predicts citation rates of ecological research[J]. Plos One, 2016, 11(11):e0166570.
[24] ZOLLER D, DOERFEL S, JÄSCHKE R, et al. Posted, visited, exported: Altmetrics in the social tagging system bibsonomy[J]. Journal of informetrics,2016,10(3):732 –749.
[25] THELWALL M, NEVILL T. Could scientists use Altmetric. com scores to predict longer term citation counts? [J]. Journal of informetrics,2018,12(1):237 –248.
[26] WANG D , SONG C , Barabasi A L . Quantifying long-term scientific Impact[J]. Science, 2013, 342(6154):127 –132.
[27] ONODERA N, YOSHIKANE F. Factors affecting citation rates of research articles[J]. Journal of the Association for Information Science and Technology,2015,66(4):739 –764.
[28] 余厚强,邱均平. 论替代计量学在图书馆文献服务中的应用[J]. 情报杂志, 2014(9):163 –166.
[30] ABRAMO G, D ANGELO, FELICI G. Predicting publication long-term impact through a combination of early citations and journal impact factor[J]. Journal of informetrics,2019,13(1):32 –49.
[31] STEGEHUIS C, LITVAK N, WALTMAN L. Predicting the long-term citation impact of recent publications[J]. Journal of informetrics,2015, 9(3):642 –657.
[32] Newman M E J. The first-mover advantage in scientific publication [J]. Europhysics letters,2009,86(6):68001.
[33] Newman M E J. Prediction of highly cited papers[J]. Europhysics letters,2014,105(2):28002.
[34] WANG M, WANG Z, CHEN G. Which can better predict the future success of articles? Bibliometric indices or alternative metrics [J]. Scientometrics,2019,119(3):1575 –1595.
[35] BHAT H S, HUANG L H, RODRIGUEZ S, et al. Citation prediction using diverse features[C]//IEEE international conference on data mining workshop,USA: IEEE, 2015:589 –596.
[36] CAO X, CHEN Y, RAY LIU K J. A data analytic approach to quantifying scientific impact[J]. Journal of informetrics,2016,10(2):471 –484.
[37] ABRISHAMI A, ALIAKBARY S. Predicting citation counts based on deep neural network learning techniques[J]. Journal of informetrics,2019,13(2): 485 –499.
[38] 吴智勇. 学术论文排序预测算法研究[D]. 内蒙古:内蒙古大学,2015.
[39] POBIEDINA N, ICHISE R. Citation count prediction as a link prediction problem[J]. Applied intelligence, 2016, 44(2): 252 –268.
[40] CHEN C. Predictive effects of structural variation on citation counts [J]. Journal of the Association for information science and technology,2014,63(3):431 –449.
[41] 于志涛,牟晓青. 文献科学计量陈氏预测指标及其应用述评[J]. 图书馆论坛,2013,33(4):32 –41.
[42] 白晓梅. 基于社会网络分析的学术影响力评估与预测[D]. 大连:大连理工大学,2017.
[43] SAYYADI H, GETOOR L. FutureRank: ranking scientific articles by predicting their future pagerank[C]//Proceedings of the SIAM international conference on data mining, USA: Society for industrial and applied mathematics,2009:533 –544.
[44] 刘大有,薛锐青,齐红. 基于作者权威值的论文价值预测算法[J]. 自动化学报,2012,38(10):1654 –1662.
[45] 樊玮,韩佳宁,张宇翔. 基于网络表示学习的论文影响力预测算法[J/OL]. 计算机工程. [2019 –06 –15]. https://doi. org/10. 19678/j. issn. 1000-3428. 0053395.
[46] 中国人工智能学会. 机器学习白皮书. [EB/OL]. [2019 –05 –20]. http://www. caai. cn/index. php? s =/home/article/detail/id/49. html.
[47] 沈雷. 基于学术网络的新论文影响力预测[D]. 济南:山东大学,2018.
[48] WANG S, XIE S, ZHANG X, et al. Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement[J]. ACM transactions on intelligent systems and technology,2016,7(4):1 –28.
[49] 曾玮. 文献排名预测算法及作者影响力评估算法研究[D]. 重庆:西南大学,2014.
[50] 杜建,武夷山. "睡美人"文献的重要特征、预测线索与政策启示[J]. 科学学研究,2018,36(11):1938 –1945.

作者贡献说明:
夏琬钧:拟定论文框架,收集资料与撰写论文;
陈晓红:论文框架指导,修订论文;
江艳萍:修订论文。

Research on Academic Paper Citation Prediction

Xia Wanjun[1,2]　Chen Xiaohong[1]　Jiang Yanping[1]

[1] Library of Southwest Jiaotong University , Chengdu 611756

[2] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756

Abstract: [**Purpose/significance**] This paper summarizes the influencing factors and prediction methods of academic paper citation, analyzes the existing problems and proposes the future development directions. [**Method/process**] This paper used the literature research method to review the research progress of academic papers at home and abroad, and summarized the relevant content and characteristics of influencing factors and prediction methods. [**Result/conclusion**] There are many indicators of influencing factors, but there is no unified selection criteria. The theoretical basis of prediction methods is weak. The research on dynamics of citation prediction is insufficient. The generality of prediction models is limited. In the future, we should strengthen the theoretical research of citation prediction methods, the combination of traditional bibliometrics and alternative metrics, the deep application of natural language processing, and establish a unified baseline standard, a more accurate prediction model.

Keywords: citation prediction　influencing factor　prediction method